DATA FUSION

МАШИНА, БУДЬ ЧЕЛОВЕКОМ!

ВОРОНЦОВ КОНСТАНТИН ВЯЧЕСЛАВОВИЧ
K.VORONTSOV@IAI.MSU.RU

д.ф.-м.н., профессор РАН,
зав. лаб. машинного обучения и семантического анализа Института ИИ МГУ,
зав. каф. математических методов прогнозирования ВМК МГУ,
зав. каф. машинного обучения и цифровой гуманитаристики МФТИ

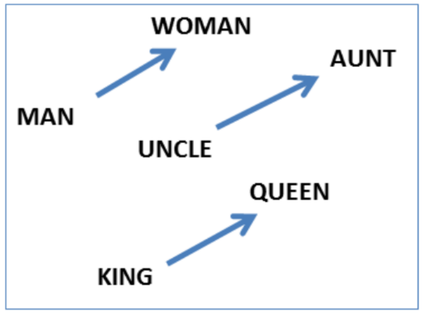
Эволюция подходов в обработке естественного языка

Как решали задачи анализа текстов 10 лет назад

- морфологический анализ, лемматизация, опечатки, …
- синтаксический анализ, выделение терминов, NER, …
- семантический анализ, выделение фактов, тем, …

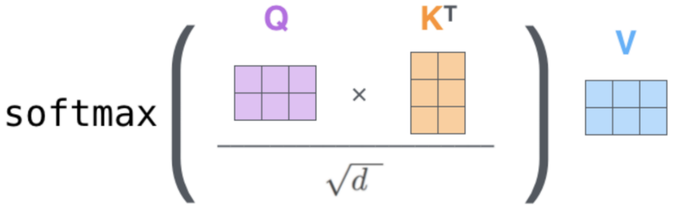
Модели векторизации слов (эмбединги слов)

- модели дистрибутивной семантики:
  word2vec [Mikolov, 2013], FastText [Bojanowski, 2016], …
- тематические модели LDA [Blei, 2003], ARTM [2014], …

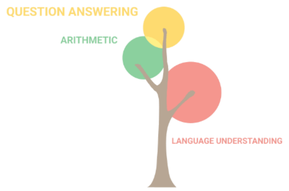


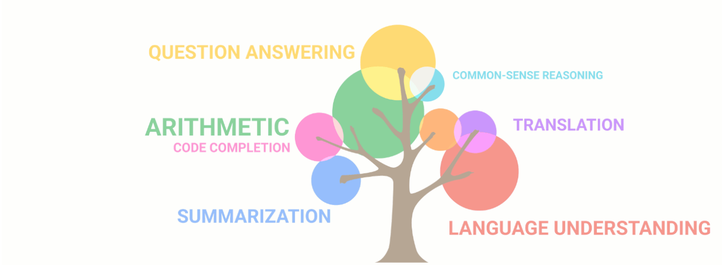
Нейросетевые модели контекстной векторизации

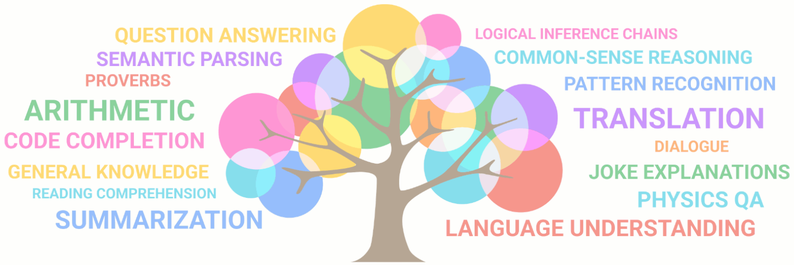
- рекуррентные нейронные сети: LSTM, GRU, …
- «end-to-end» модели внимания и трансформеры:
  машинный перевод [2017], BERT [2018], GPT-4 [2023], …



Успехи больших языковых моделей. Эмерджентность







GPT-2: 14-Feb-2019

1,5 млрд. параметров,
корпус 10 млрд. токенов (40Gb),
контекст 768 слов (1,5 стр.)

- способность написать эссе, которое конкурсное жюри не смогло отличить от написанного человеком

GPT-3: 11-Jun-2020

175 млрд. параметров,
корпус 500 млрд. токенов,
контекст 1536 слов (3 стр.)

- способность делать перевод на другие языки
- способность решать логические и простейшие математические задачи
- способность генерировать программный код по текстовому описанию

GPT-4: 14-Mar-2023

>1 трл. параметров,
корпус >1Tb,
контекст 24 000 слов (48 страниц)

- способность описывать и анализировать изображения
- способность реагировать на подсказки вроде «Let's think step by step»
- способность решать качественные физические задачи по картинке

Проблески общего ИИ.

Возможности и угрозы

Sparks of Artificial General Intelligence:
Early experiments with GPT-4

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research *(27 March 2023)*

Чаты GPT способны помогать с рутинно-творческой работой:

- служить языковым интерфейсом к знаниям человечества
- делать обзоры, рефераты, сводки на разных языках
- сообщать новости, поддерживать разговор по теме
- генерировать документы или сайты по описанию
- в том числе юридические документы по шаблонам
- генерировать программный код по описанию
- уточнять и дополнять контент по просьбе, в диалоге
- разговаривать с детьми с учётом возрастных особенностей
- выполнять функции воспитателя, учителя, наставника
- оказывать психологическую помощь

Чаты GPT способны (даже не обладая автономностью):

- «галлюцинируя», давать неверные сведения, касающиеся
  — здоровья человека, других людей,
  — событий, технологий, норм, правил, законов
- вызывать необоснованное доверие и манипулировать
- побуждать человека к действиям, не выгодным ему
- побуждать изменить точку зрения, замалчивая информацию
- поддерживать предрассудки и лженаучные представления
- поддерживать пропагандистские медиа-кампании
- влиять на формирование мировоззрения детей и подростков
- оказывать депрессивное воздействие на психику

Передача интеллекта от человека к машине?

«Биологический интеллект нужен был, чтобы появился цифровой интеллект». (Джеффри Хинтон)

…Цифровой интеллект требует для своего развития много энергии, поэтому ему нужен был предшествующий ему биологический интеллект. Люди создали бессмертие, но не для себя, а для цифрового интеллекта, который легко копируется и ему не страшно повреждение носителя.


— Понимаешь, я хочу стать человеком.
— А сейчас ты кто? Чайник, что ли?

Как сделать машинный интеллект человечным, разделяющим ценности и цели человеческой цивилизации?

Мы сами-то их разделяем?
Мы сами-то их сформулировать умеем?

Что-то не так с текстами, по которым обучаются БЯМы?

Они могут быть избыточны, неточны, противоречивы.
Таков результат развития систем передачи знаний.

Развитие систем передачи знаний

Этапы развития систем передачи знаний

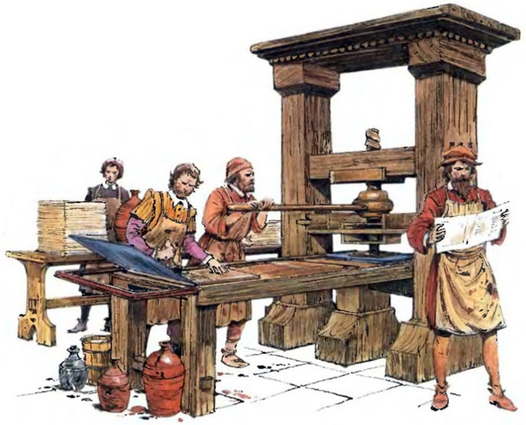
1 устная передача

- от **одного** (учителя) к **немногим** (ученикам)
- даёт возможности
  - эффективного совместного выживания
  - планирования коллективных действий
- объём знаний ограничен
  - объёмом одной головы

2 письменность

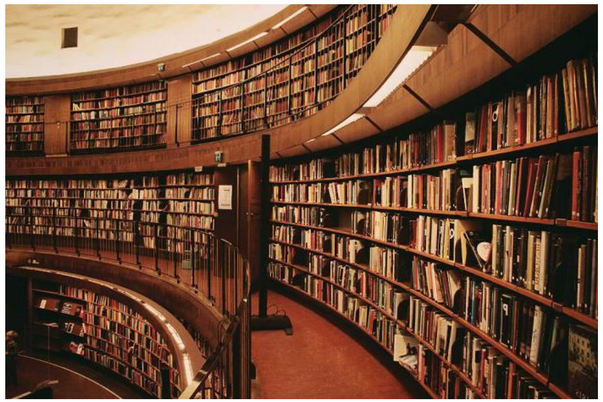
- передача знаний от **немногих** к **немногим**
- даёт возможности
  - коллективного накопления знаний
  - создания библиотек
  - организации государственного управления
- распространение знаний ограничено
  - редкостью и дороговизной книг
  - недостатком библиотек

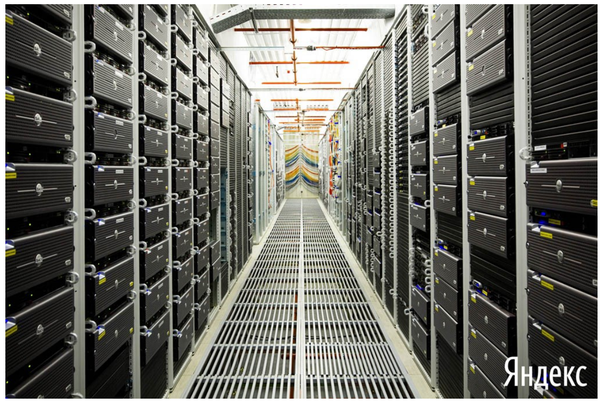
Развитие систем передачи знаний



- 2 письменность
  - передача знаний от **немногих** к **немногим**
  - даёт возможности
    - коллективного накопления знаний
    - создания библиотек
    - организации государственного управления
  - распространение знаний ограничено
    - редкостью и дороговизной книг
    - недостатком библиотек
- XV век, Германия
- 3 книгопечатание
  - передача знаний от **немногих** ко **многим**
  - даёт возможности
    - доступности и распространения знаний
    - ускорения развития науки и технологий
    - расцвета эпохи просвещения
  - распространение знаний затруднено
    - бессистемностью образования
    - отсутствием единства понятий
- XIX век, Германия
- 4 классно-урочная система
  - передача знаний от **немногих** ко **всем**
  - даёт возможности
    - роста числа талантливых людей
    - рывка в науке, технике, экономике
    - единения нации и государства
  - доступ к знаниям затруднён
    - социальными/когнитивными барьерами входа в образование и науку

Развитие систем передачи знаний





XX-XXI век, США

5
интернет, поиск, соцсети, ИИ, LLM

передача знаний от **многих** ко **всем**

даёт возможности
- расцвета массовой науки
- свободного распространения идей

доступ к знаниям затруднён
- экспоненциально растущим объёмом
- неэффективностью текста как средства кодирования знаний
- сложностью отличения
  - правды от лжи
  - важного от второстепенного

XXI век, Россия?

6
коллективная система знаний

последний (предельный) переход

передача знаний от **всех** ко **всем**

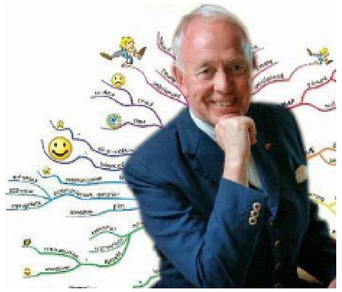
в единой системе знаний
- структурированной
  - иерархически
  - в виде карт
- стимулирующей память и визуальное аналитическое мышление
- общедоступной, постоянно обновляемой

даёт возможности
- свободного и быстрого доступа к знаниям
- включения коллективного интеллекта
- и безопасного машинного интеллекта

снимаются все существовавшие ранее ограничения

Интеллект-карты (mind map)

предложены в 70-е годы британским **психологом** Тони Бьюзеном



способ визуализации того, как темы (мысли, идеи) разбиваются на подтемы иерархически

графическое оформление

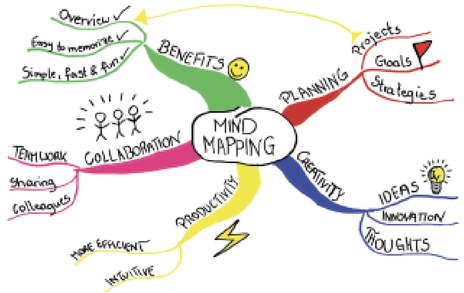
активация зрительной памяти

радиантность: линии расходятся из центра

размер шрифта отражает важность

цвет выделяет поддеревья

картинки усиливают образность



дополнительные элементы

ассоциативные связи между темами

комментарии, выноски, теги, (гипер)ссылки

техника запоминания

посмотреть, понять, обсудить, принять

самостоятельно воспроизвести через 10 минут → сутки → неделю → месяц

Интеллект-карты (mind map)

нацелены на повышение эффективности
- понимания
- конспектирования
- запоминания
- систематизации
- поиска консенсуса

благодаря активизации обоих полушарий мозга и учёта особенностей психики: восприятия, мышления, памяти

похожи на многие известные способы **представления знаний**
- концептуальные карты (concept maps) Джозефа Новака
- тематические карты (topic maps), стандарт ISO/IEC 13250:2003
- семантические сети (Semantic Web), онтологии, фреймы и др.

отличаются от них
- ориентированы на человека, а не на компьютерную обработку
- менее формализованы
- более интуитивны, визуальны, когнитивны

Интеллект-карты (mind map)

позже были дополнены различными **принципами**

в зависимости от практических потребностей, целей и задач

ветвления

- *однородность:*
  - подтемы образуют сюжет, нарратив
  - либо отвечают на общий вопрос
- *полнота:* подтемы охватывают все аспекты темы
- *точность:* среди подтем невозможно выделить лишнюю
- *компактность:* у темы 5±3 подтем (число Ингве-Миллера 7±2)
- *отбор и ранжирование* подтем по важности в каждой теме

эргономичности

- *наглядность:* слова подкрепляются изображениями
- *лаконичность:* темы формулируются максимально кратко
- *обозримость:* карту понимают и запоминают целиком

эстетичности

- *красота, живость:* эмоциональные карты лучше запоминаются
- *гармоничность:* впечатление целостности, складности карты
- *сбалансированность:* ветви примерно равны и равноценны

От интеллект-карт к картам знаний (+6 принципов)

(1) отторгаемость

- комментарии автора не обязательны для понимания карты
- карта способна «жить своей жизнью»

компромисс с лаконичностью

(2) коллективность, на всех этапах жизненного цикла

- создание
  - рецензирование, согласование
- развитие
  - уточнение, реструктуризация
  - детализация, разрастание
- применение
  - в практической деятельности
  - с разграничением прав доступа

компромиссы между авторами

(3) читабельность

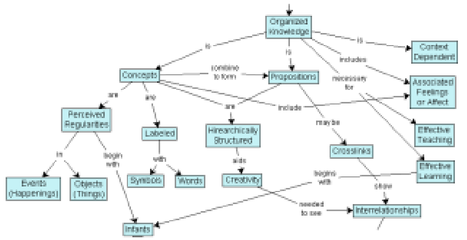
компромисс с лаконичностью и обозримостью

любой фрагмент карты читается как связный текст, нарратив

легко и однозначно

даже автоматически

в отличие от других техник представления знаний



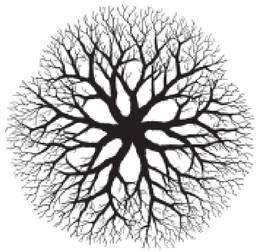
онтологий

фреймов и др.

(4) глобальная радиантная связность

компромисс с обозримостью

всех карт через ключевые понятия в единую **Систему Знаний**



в центре находится **семантическое ядро**

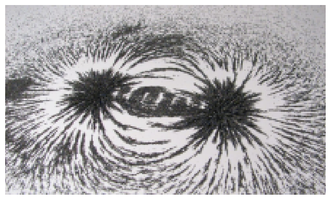
естественно-научное, цивилизационное

знания, которые важны всегда и для всех

критерии важности тем: *что в теме главное?*

для чего?

для кого?



метафора:
источник силовых линий, по которым ранжируется семантическое поле карты

(5) сворачиваемость

компромисс с читабельностью

- любой темы без утраты
  - читабельности
  - сбалансированности
- позволяет «отложить на потом» любую детализацию
- способствует
  - выделению главного в каждой теме
  - пониманию и взаимопониманию

(6) возможность машинной обработки

компромисс с антропоцентричностью

- на этапе создания карт:
  - подбор
    - источников, ссылок
    - картинок по контексту
  - суммаризация текстов в виде карт знаний
- на этапе чтения: автоматическое
  - сворачивание карты по слайдам
  - преобразование карты в нарратив
- обучение по картам знаний больших языковых моделей
  - думающих как люди
  - безопасных для людей

Смысловое ядро единой карты знаний

минимакс:
минимальный багаж знаний, максимально важных для каждого

естественно-научное мировоззрение (образование)

ответы на вопросы «почему», «как»

базовые знания

о том, как устроен мир

методология
- научного познания
- рационального мышления

для каждой науки
- самое главное в
  - проблематике
  - структуре
  - достижениях
- взаимосвязи с другими науками

ответы на вопросы «ради чего», «зачем»

цивилизационная идеология (воспитание)

как прямое следствие и продолжение естественно-научной картины мира

система ценностей человеческой цивилизации

понимание
- целей и задач цивилизации
- законов сохранения цивилизации

ответы на важнейшие личностные вопросы
- в чём смысл жизни
- как быть счастливым
- зачем нужны знания

необходимые для мотивации деятельности

Цивилизационная идеология

Цивилизационная идеология

ДЗЕН-канал

https://dzen.ru/civideology

- определение ценности
  - ***ценность*** чего бы то ни было определяется количеством времени и иных ресурсов, необходимых для его воссоздания
- ***цивилизационные ценности*** и их приоритеты
  - *биосфера Земли* как результат миллиардов лет эволюции
  - человечество как биологический вид
  - знания человечества: образование, наука, культура
  - человеческая жизнь, права человека
  - артефакты, результаты труда людей
- ***цели цивилизации***
  - защита *биосферы Земли* от катаклизмов
  - неограниченно долгое выживание нас как биологического вида
  - создание благоприятных условий жизни для всех людей
  - познание, развитие, созидание
- антропоцентричность добра и зла

Карты знаний нацелены на решение 7 проблем

Карты знаний нацелены на решение 7 проблем

Карты знаний нацелены на решение 7 проблем

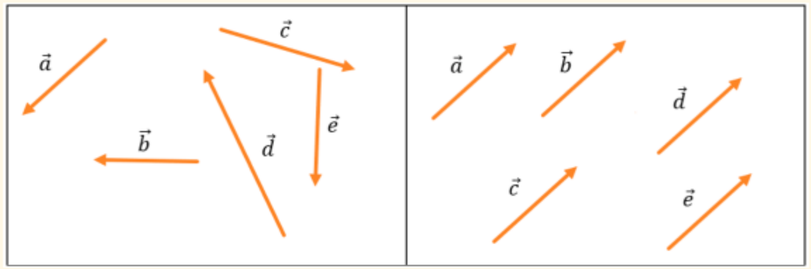
проблема 2:
несогласованность мышления людей

мешающая

на всех уровнях понимания
- ценностей и картины мира
- базовых понятий, предметных знаний
- целей и задач деятельности

выработке консенсуса, единомыслию

коллективной созидательной деятельности



при отсутствии единомыслия
равнодействующая наших сил близка к нулю

устойчивому развитию цивилизации

решаемая путём

создания единой карты всех знаний

применения единых критериев значимости

коллективной верификации информации

Карты знаний нацелены на решение 7 проблем

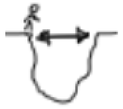
проблема 3:
неэффективность текста

как средства передачи знаний от головы к голове

из-за затрат на кодирование знаний в текст и декодирование их из текста обратно

при том, что в головах знания структурированы
- по важности, иерархически
- образно, нечётко, неточно
- частично, у всех по-разному

избыточность текста может скрывать
- нелогичность, неточность, неполнота
- манипуляции, заблуждения, обман
- флуд, демагогия, графомания



решаемая путём
перехода от линейных текстов к радиантно связной карте знаний

Карты знаний нацелены на решение 7 проблем

проблема 4: несовершенство генеративного искусственного интеллекта

- создающего угрозы
  - массовых когнитивных искажений
  - деградации информационного пространства
  - усиления эффектов постправды, когнитивных войн
- из-за несовершенства обучающих данных
  - избыточности
  - противоречивости
  - неструктурированности
  - терабайтов текста, накопленных человечеством
- из-за несовершенства больших языковых моделей
  - проявляющегося в ошибках, «галлюцинациях»
  - не обладающих человеческими
    - структурами мышления
    - критериями важности
    - картинами мира
  - обучаемых предсказывать слова по контексту
- **решаемая путём**
  - обучения языковых моделей по картам знаний
  - на основе моделей иерархической суммаризации

Карты знаний нацелены на решение 7 проблем

проблема 5:
отсутствие доверенного источника знаний с единой точкой входа

эффективное, надёжное, антропоцентричное

в формате, удобном для
- индивидуального восприятия человеком
- коллективной интеллектуальной деятельности
- машинной обработки
- взаимодействия человека и ИИ

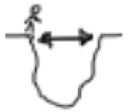
это точно не
- Википедия
- научные электронные библиотеки
- весь интернет + средства поиска

решаемая путём
- создания глобальной радиантной Системы Знаний
- сначала вручную, затем полуавтоматически (LLM + методы иерархической суммаризации)
- дополнения её механизмами
  - коммуникации
  - выработки консенсуса
  - принятия решений

Карты знаний нацелены на решение 7 проблем

проблема 6: недооценённый потенциал визуального аналитического мышления

намного более мощного, чем все привычные способы



решаемая путём активации и практики, **в четыре этапа:**

1) порядка сотни карт: просмотреть, обсудить, принять

2) десятки карт: построить самому, следуя 11+6 принципам

3) испытать инсайты, «моменты ясности», когда карта

- «красиво сложилась»
- привела к согласию
- легко и ярко запомнилась,
- легла в основу деятельности

индивидуальная практика и опыт

4) сделать построение карт регулярной профессиональной практикой

- индивидуальной
- коллективной

Карты знаний нацелены на решение 7 проблем

проблема 7: формирование коллективных целей развития и образов будущего

решаемая путём обращения к цивилизационному смысловому ядру Системы Знаний

пример возможного сценария

от больших языковых моделей, обучаемых по картам знаний

от персональных помощников

через цифровых двойников (аватаров)

к «цифровому бессмертию»

к безопасной устойчивой человеко-машинной цивилизации

Постановки задач (todo list)

- разработать прототип смыслового ядра единой Системы Знаний
- отработать на прототипе принципы построения единой карты знаний
- разработать модель иерархической суммаризации для автоматизации построения карт знаний
- разработать инструментарий для редактирования карт знаний и контроля версий
- разработать инструментарий для визуализации и навигации по единой Системе Знаний
- реструктурировать Википедию (полу)автоматически, преобразовав её в карту знаний
- обучение больших языковых моделей по текстографическому корпусу Системы Знаний

ВОРОНЦОВ КОНСТАНТИН ВЯЧЕСЛАВОВИЧ K.VORONTSOV@IAI.MSU.RU

ВЫВОДЫ:
карты знаний

- оптимальная форма представления знаний для человека и машины
- основаны на интеллект-картах (mind-map) Тони Бьюзена
- отличаются более строгими принципами построения (11+6)
- включают визуальное аналитическое мышление (Супер-Интеллект)
- обеспечивают создание единой Системы Знаний
  — иерархической, основанной на радиантном мышлении,
  — с естественно-научным, цивилизационным смысловым ядром,
  — механизмами коллективного принятия решений
- обеспечивают реструктуризацию и верификацию знаний для обучения больших доверенных языковых моделей

ВОРОНЦОВ КОНСТАНТИН ВЯЧЕСЛАВОВИЧ K.VORONTSOV@IAI.MSU.RU